2022-2023 National Survey of Family Growth: Variance Estimation

Prepared Under Contract #75D30120C09732 with

National Center for Health Statistics

3311 Toledo Road

Hyattsville, MD 20782

Prepared by:

Taylor Lewis, Stephanie Zimmer, Jennifer Cooney, Andy Peytchev

RTI International

3040 East Cornwallis Road

Research Triangle Park, NC 27709-2194

August 2025

Table of Contents

1. Background

This report describes the methods used to convert the original National Survey of Family Growth (NSFG) sample error design indicators (i.e., stratum and cluster codes) into masked versions that enable variance estimation of point estimates and other analytic quantities using software capable of analyzing complex survey data, such as SUDAAN (RTI International, 2012), the SURVEY procedures in SAS® (Lewis, 2017), the svy commands in Stata (Heeringa et al., 2017) and the srvyr() and survey() packages in R (Zimmer et al., 2025). The methods described below were applied to the indicators associated with the first 2 years of the NSFG's continuous 8-year design, the 8 quarters of data collection between January 2022 and December 2023. More details on the broader NSFG sample design can be found in the **Sample Design** report.

Individuals are selected for the NSFG main survey by way of a multistage, hierarchically clustered sample design. In the first stage, a sample of primary sampling units (PSUs) comprising individual counties or groupings of counties is selected. In the second stage, a sample of secondary sampling units (SSUs) defined as Census Block Groups or groupings thereof within PSUs is selected. In the third stage, a sample of addresses (a proxy for households) within the sampled SSUs is selected. Each of these three stages can be considered cluster samples. In the fourth stage, one individual is selected from a household's roster and asked to participate in the main survey.

There are several reasons that stratum and cluster identifiers in a multistage sample survey such as NSFG require collapsing or recoding. The primary reason is to disguise geographic information for disclosure avoidance purposes, to protect NSFG respondents' identities and associated responses to the survey. A second reason is to simplify variance estimation calculations. One common tool for simplifying variance computations is to invoke the *ultimate cluster assumption* (Kish, 1965; Kalton, 1983), which treats PSUs as selected with replacement, even though they are typically selected without replacement. Under the ultimate cluster assumption, variance estimation reduces to a function of PSU-level totals, meaning variance components corresponding to subsequent sampling stages need not be explicitly computed. Some survey software (e.g., SAS) effectively requires analysts of multistage cluster designs to adopt the ultimate cluster assumption, whereas other software such as SUDAAN and Stata permits users to explicitly account for each stage of sampling. Although the impact of adopting the ultimate cluster assumption can vary across sample designs, it most often results in a slight overestimation of variance. This tendency can be seen in prior empirical comparisons using prior data from the NSFG (West, 2010) and the National Ambulatory Medical Care Survey (Hing et al., 2003).

In general, these collapsing and/or recoding techniques result in "masked" pseudo-strata and pseudo-PSUs. At a minimum, each pseudo-stratum must contain two or more pseudo-PSUs. More PSUs per stratum can increase the reliability of variance estimates and will increase the degrees of freedom for inferences. On the other hand, reducing the number of strata by collapsing will tend to result in over-estimates of sampling variance. These losses in precision are variable-specific but tend to be small as most of the gains from stratification occur with few strata (Cochran, 1968). Losses would only be substantive in scenarios where the variable is highly correlated with the stratification variable(s), which is less likely in the case of NSFG considering its geographic stratification scheme. For the 2022-2023 NSFG data collection period, strata from the original design were collapsed, but this was done in a way that would minimize losses. Specifically, strata that were expected to be most similar to each other using the design information available from the sampling frame were collapsed together.

2. Methods

During the 2022-2023 NSFG data collection period, a total of 979 unique SSUs were released from 65 unique PSUs: (1) 25 of these were non-certainty PSUs allocated to one of the two years (50 total), and (2) 15 of these were certainty PSUs allocated to both years. With few exceptions, 3 SSUs were sampled from each PSU in each quarter, for 12 SSUs per year per PSU. A three-pronged approach was used to create a total of 80 variance estimation clusters (VECLs) (i.e., pseudo-PSUs) nested within 20 variance estimation strata (VESTs) (i.e., pseudo-strata), resulting in 4 unique VECLs per VEST.

The approach taken was as follows:

1. In the largest certainty PSU, the 48 SSUs were randomly divided into four sets of 12 SSUs, each of which defines its own VECL code. These are all maintained within the same VEST code. In all, this PSU contributes four VECL codes within one VEST code.
2. In the remaining 14 certainty PSUs, a total of 24 SSUs were sampled across the 2 years. Each PSU's set of 24 SSUs was first randomly divided into two sets of 12, resulting in two VECLs per PSU. Thus, these 14 PSUs produced a total of 28 VECLs. The two VECLs for a given PSU were paired with another PSU to establish four VECLs within a single VEST code. In all, these 14 certainty PSUs contribute 28 VECL codes within 7 VEST codes.
3. For the 50 non-certainty PSUs, the 12 SSUs sampled each year were collapsed into a single VECL code. These were then combined with the VECL codes of three other PSUs within the same design stratum to create a set of four VECLs within a single VEST code. On a few occasions, the SSUs for more than one PSU were combined into a single VECL code, and occasionally across design stratum boundaries. In all, these 50 non-certainty PSUs contributed 48 VECL codes within 12 VEST codes.

Once all VECL and VEST codes were created, the four VECLs were arbitrarily coded 1, 2, 3, or 4, and the VEST codes were scrambled and assigned labels 1, 2, …, 20 at random.

In expectation, this procedure results in a modest increase in the magnitude of variance estimates relative to the original design (i.e., using the original stratum and cluster identifiers). Table 1 demonstrates this via a side-by-side analysis of weighted key estimates and standard errors, along with the design effect (Kish, 1965) of the given estimate. One can note how most, but not all, of the standard errors are larger under the masked design.

Table 1: Comparison of point estimates for masked variance estimation method (with VEST/VECL codes) and the original sample design stratum and cluster codes

	Female					
	Masked Design			Original Design		
Outcomes	**Percent**	**SE Percent**	**Design Effect**	**Percent**	**SE Percent**	**Design Effect**
Age at first sex (<15)	14.5	0.77	2.20	14.5	0.70	1.77
Age at first sex (15-17)	39.6	1.20	2.73	29.6	1.09	2.25
Age at first sex (18+)	45.9	1.50	4.10	45.9	1.33	3.24
Ever cohabited	51.0	1.00	2.22	51.0	1.08	2.61
No live births	50.5	1.15	2.97	50.5	1.14	2.92
One live birth	15.3	0.58	1.51	15.3	0.57	1.41
Two or more live births	34.2	1.08	2.86	34.2	1.06	2.81
Intend a/another birth	46.2	0.83	1.51	46.2	0.89	1.74
Used contraception at first sex	68.9	1.05	2.30	68.9	1.03	2.20
Had sex in the last 12 months	83.7	0.64	1.36	83.7	0.66	1.43
Ever smoked at least 100 cigarettes	20.0	0.93	3.00	20.0	0.88	2.68
Ever had an HIV test outside of blood donation	43.1	1.18	3.16	43.1	1.09	2.68
Health care coverage in last 12 months	89.5	0.71	2.89	89.5	0.65	2.45
Received public assistance in the last 12 months	6.0	0.61	3.34	6.0	0.54	2.72
Ever pregnant	54.7	1.16	3.02	54.7	1.12	2.80

	Male					
	Masked Design			Original Design		
Outcomes	**Percent**	**SE Percent**	**Design Effect**	**Percent**	**SE Percent**	**Design Effect**
Age at first sex (<15)	18.1	1.10	2.12	18.1	1.07	2.05
Age at first sex (15-17)	42.7	1.08	1.26	42.7	1.08	1.26
Age at first sex (18+)	39.2	1.39	2.14	39.2	1.37	2.07
Ever cohabited	27.7	0.96	2.00	27.7	0.94	1.95
No biological children	50.0	1.22	1.97	50.0	1.32	2.32
One biological child	16.1	0.75	1.40	16.1	0.80	1.58
Two or more biological children	33.7	1.31	2.49	33.7	1.25	2.33
Intend a/another birth	55.1	0.88	1.32	55.1	1.04	1.85
Used contraception at first sex	74.1	1.29	2.28	74.1	1.23	2.05
Had sex in the last 12 months	84.7	0.75	1.44	84.7	0.77	1.53
Ever smoked at least 100 cigarettes	28.5	1.13	2.74	28.5	1.06	2.37
Ever had an HIV test outside of blood donation	33.3	1.08	2.29	33.3	1.14	2.53
Health care coverage in last 12 months	85.2	0.93	2.89	85.2	0.85	2.45
Received public assistance in the last 12 months	5.1	0.45	1.71	5.1	0.50	2.10

3. Software Considerations

Data users are reminded that standard statistical procedures and software generally assume that data are generated via simple random sampling and will tend to produce incorrect estimates of variances and standard errors when used to analyze data from the NSFG. Specifically, variance estimates and standard errors would likely be, on average, too small, and therefore yield results subject to excessive Type I error. Analysts should use appropriate software to account for the NSFG's complex sample design in their analyses. Several software packages are available for analyzing complex samples, including Stata, SAS, R, Stata, and SUDAAN. The key design variables for analysis of 2022-2023 NSFG data are:

- VEST: Stratum variable
- VECL: Cluster variable
- WGT2022_2023: Final analysis weight

Examples of program statements in SAS and Stata that illustrate the correct use of the design variables for variance estimation can be found on the webpage for the 2022-2023 public use release under the title "Variance Estimation Examples." Examples of program statements in R and SUDAAN can be found in the **Study Design and Data Collection Procedures** report.

References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295–313.

Heeringa, S., West, B., & Berglund, P. (2017). *Applied survey data analysis*. 2nd edition. Boca Raton, FL: Chapman & Hall/CRC Press.

Hing, E., Gousen, S., Shimizu, I., & Burt, C. (2003). Guide to using masked design variables to estimate standard errors in public use files of the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey. *Inquiry, 40*, 401-415.

Kalton, G. (1983). *Introduction to survey sampling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-035. Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Lewis, T. (2017). *Complex survey data analysis with SAS®*. Boca Raton, FL: Chapman & Hall/CRC Press.

RTI International. (2012). *SUDAAN: Statistical software for weighting, imputing, and analyzing data, Release 11*. Research Triangle Park, NC: Research Triangle Institute.

West, B. (2010). *Accounting for multi-stage sample designs in complex sample variance estimation*. Institute for Social Research Technical Report Prepared for National Survey of Family Growth (NSFG) User Documentation. Available at https://smponline.isr.umich.edu/asda/first_stage_ve_new.pdf.

Zimmer, S., Powell, R., & Velásquez, I. (2025). *Exploring complex survey data analysis using R: A tidy introduction with {srvyr} and {survey}*. Boca Raton, FL: Chapman & Hall/CRC Press.